

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-075995

(43)Date of publication of application : 18.03.1994

(51)Int.Cl.

G06F 15/40
G06F 15/401

(21)Application number : 04-250385

(71)Applicant : OMRON CORP

(22)Date of filing : 27.08.1992

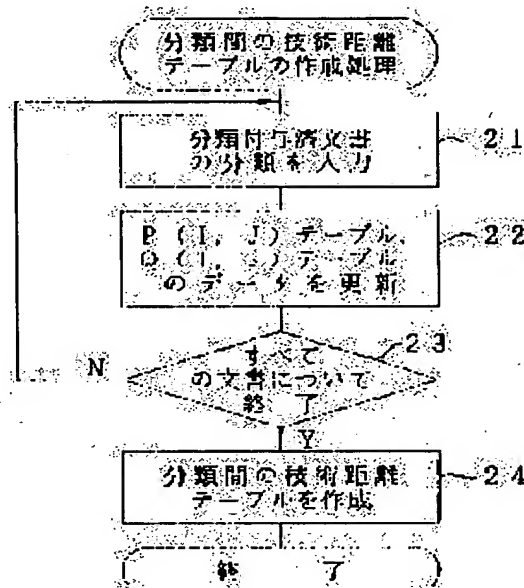
(72)Inventor : KISHIOOJI YASUAKI
OZAKI TOKIO
KUNO ATSUSHI

(54) AUTOMATIC DEVICE AND METHOD FOR IMPARTING CLASSIFICATION

(57)Abstract:

PURPOSE: To attain the imparting of a correct classification based on a data base by automatically generating the data base used for imparting the classification.

CONSTITUTION: Based on the classification and keyword of a classification imparted document to which the classification is already imparted, a keyword/ classification table is generated, which stores the classification in close relation with the keyword and the degree indicating the depth of classification relation at every inter-classification distance table P, Q and keyword showing the strength of relation between two classifications. When the plural keywords included in a classification nonimparted document are inputted, the keyword/ classification table is referred, the total value of the relation degree of the classification relating the inputted keyword is calculated at every classification and the candidate of classification to be imparted is selected according to the order of the largeness of the total value. It is inspected whether or not mutual distance between plural selected candidate classifications is within a proper range by referring to the inter-classification distance tables P and Q and the candidate classification is decided to be the final classification when it is within the proper range.



BEST AVAILABLE COPY

LEGAL STATUS

[Date of request for examination]

16.03.1999

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

3333998

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平6-75995

(43)公開日 平成6年(1994)3月18日

(51)IntCl.⁵

G 0 6 F 15/40
15/401

識別記号

5 0 0 T

庁内整理番号

7218-5L

7218-5L

F I

技術表示箇所

審査請求 未請求 請求項の数9(全14頁)

(21)出願番号 特願平4-250385

(22)出願日 平成4年(1992)8月27日

(71)出願人 000002945

オムロン株式会社

京都府京都市右京区花園土堂町10番地

(72)発明者 岸大路 泰明

京都府京都市右京区花園土堂町10番地 オムロン株式会社内

(72)発明者 尾崎 時夫

京都府京都市右京区花園土堂町10番地 オムロン株式会社内

(72)発明者 久野 敦司

京都府京都市右京区花園土堂町10番地 オムロン株式会社内

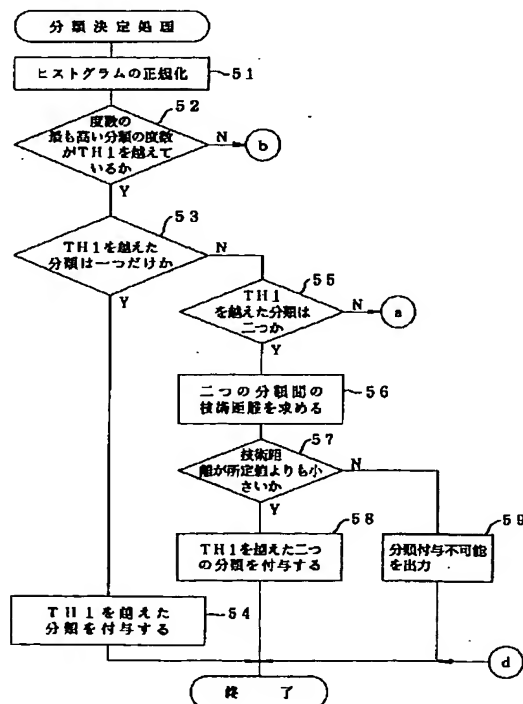
(74)代理人 弁理士 牛久 健司

(54)【発明の名称】 自動分類付与装置および方法

(57)【要約】

【目的】 分類付与のために用いるデータ・ベースを自動的に作成し、このデータ・ベースに基づいて正しい分類の付与を達成する。

【構成】 既に分類が付与された分類付与済文書の分類およびキーワードに基づいて、2つの分類間の関連性の強さを表わす分類間距離テーブル、ならびにキーワードごとに、そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合を記憶したキーワード／分類テーブルが作成される。分類未付与文書に含まれる複数のキーワードが入力されると、キーワード／分類テーブルを参照して、入力されたキーワードに関連する分類の関連度合の合計値が分類ごとに算出され、この合計値の大きさの順序にしたがって付与すべき分類の候補が選択される。分類間距離テーブルを参照して、選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査され、妥当な範囲内にあれば候補分類が最終分類として決定される。



【特許請求の範囲】

【請求項1】 分類未付与文書に含まれる複数のキーワードを入力する手段、キーワードごとに、そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して、入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し、この合計値の大きさの順序にしたがって付与すべき分類の候補を選択する手段、ならびに2つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して、選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し、妥当な範囲内にあれば候補分類を最終分類と決定する手段、を備えた自動分類付与装置。

【請求項2】 上記決定手段は、上記合計値が所定値以上である候補分類が1つである場合に、上記分類間距離テーブルを参照することなくその候補分類を最終分類と決定する、請求項1に記載の自動分類付与装置。

【請求項3】 上記合計値が所定値よりも大きい候補分類がない場合に、上記合計値の大きさの順序にしたがって複数の分類からなる分類パターンを作成し、同一分類パターンが所定回数出現したときに新たな分類を創設して付与する手段、をさらに備えた請求項1に記載の自動分類付与装置。

【請求項4】 一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類の組を入力するための手段、および入力された分類の組に2つの分類が同時に含まれる程度に基づいて、2つの分類間の距離を、すべての分類の中から選択されたすべての組合せの分類対について求め、分類間距離テーブルを作成する手段、を備えた分類間距離テーブル作成装置。

【請求項5】 分類があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力するための手段、入力されたキーワードごとに、それらのキーワードに関連する分類の関連度合を求め、関連度合の大きさの順序にしたがって所定数の分類を選択する手段、キーワードごとに、それに関連する選択された分類の関連度合に基づいてキーワードを評価し、関連度合の低い分類のみが関連するキーワードを削除するキーワード評価手段、および削除されずに残ったキーワードのそれぞれについて、そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成する手段、を備えたキーワード／分類テーブル作成装置。

【請求項6】 上記キーワード評価手段は、2つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照してキーワードを評価し、分類間距離の大きい2つの分類が関連するキーワードを削除する

2

ものである、請求項5に記載のキーワード／分類テーブル作成装置。

【請求項7】 分類未付与文書に含まれる複数のキーワードを入力し、キーワードごとに、そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して、入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し、この合計値の大きさの順序にしたがって付与すべき分類の候補を選択し、2つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して、選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し、妥当な範囲内にあれば候補分類を最終分類と決定する、自動分類付与方法。

【請求項8】 一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類の組を入力し、入力された分類の組に2つの分類が同時に含まれる程度に基づいて、2つの分類間の距離を、すべての分類の中から選択されたすべての組合せの分類対について求め、分類間距離テーブルを作成する、分類間距離テーブル作成方法。

【請求項9】 分類があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力し、入力されたキーワードごとに、それらのキーワードに関連する分類の関連度合を求め、関連度合の大きさの順序にしたがって所定数の分類を選択し、キーワードごとに、それに関連する選択された分類の関連度合に基づいてキーワードを評価し、関連度合の低い分類のみが関連するキーワードを削除するとともに、2つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照してキーワードを評価し、分類間距離の大きい2つの分類が関連するキーワードを削除し、削除されずに残ったキーワードのそれぞれについて、そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成する、キーワード／分類テーブル作成方法。

【発明の詳細な説明】

【0001】

【技術分野】この発明は、学術論文、科学記事、特許公報とその抄録、その他の文書を複数のカテゴリーに分類する自動分類付与装置および方法、ならびにこの自動分類付与のために用いる分類間距離テーブルやキーワード／分類テーブルの作成装置および方法に関する。

【0002】

【従来技術とその問題点】従来の自動分類付与装置には、特開平1-188934号公報や特開平2-98778号公報に記載のものがある。これらの装置は電子化された文書

からキーワードを抽出しキーワードの頻度だけで分類を決定したり、あらかじめ作成された生成規則を用いるものである。しかしながら頻度だけに基づいたのでは全般的はずれな分類を付与してしまう可能性があり、また生成規則を用いるものでもあらかじめ人間が生成規則辞書を作成しないといけないという煩わしさがあった。

【0003】

【発明の開示】この発明は、既に分類が付与された文書に基づいて自動分類付与のためのデータ・ベースを作成し、このデータ・ベースに基づいて適切な分類を付与することのできる装置および方法を提供するものである。

【0004】この発明による自動分類付与装置は、分類未付与文書に含まれる複数のキーワードを入力する手段、キーワードごとに、そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して、入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し、この合計値の大きさの順序にしたがって付与すべき分類の候補を選択する手段、ならびに2つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して、選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し、妥当な範囲内にあれば候補分類を最終分類と決定する手段を備えている。

【0005】この発明の実施態様においては、上記決定手段は、上記合計値が所定値以上である候補分類が1つである場合に、上記分類間距離テーブルを参照することなくその候補分類を最終分類と決定する。

【0006】この発明の他の好ましい実施態様においては、上記合計値が所定値よりも大きい候補分類がない場合に、上記合計値の大きさの順序にしたがって複数の分類からなる分類パターンを作成し、同一分類パターンが所定回数出現したときに新たな分類を創設して付与する手段がさらに設けられる。

【0007】この発明による自動分類付与方法は、分類未付与文書に含まれる複数のキーワードを入力し、キーワードごとに、そのキーワードに関連の深い分類およびその分類の関連の深さを示す度合をあらかじめ記憶したキーワード／分類テーブルを参照して、入力されたキーワードに関連する分類の関連度合の合計値を分類ごとに算出し、この合計値の大きさの順序にしたがって付与すべき分類の候補を選択し、2つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照して、選択された複数の候補分類相互間の距離が妥当な範囲内にあるかどうかを検査し、妥当な範囲内にあれば候補分類を最終分類と決定するものである。

【0008】この発明は上述した自動分類付与装置および方法で用いられる分類間距離テーブルを作成する装置および方法を提供している。

【0009】この発明による分類間距離テーブル作成装

置は、一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類の組を入力するための手段、および入力された分類の組に2つの分類が同時に含まれる程度に基づいて、2つの分類間の距離を、すべての分類の中から選択されたすべての組合せの分類対について求め、分類間距離テーブルを作成する手段を備えている。

【0010】この発明による分類間距離テーブル作成方法は、一文書について複数の分類からなる分類の組があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類の組を入力し、入力された分類の組に2つの分類が同時に含まれる程度に基づいて、2つの分類間の距離を、すべての分類の中から選択されたすべての組合せの分類対について求め、分類間距離テーブルを作成するものである。

【0011】この発明はさらに上記自動分類付与装置および方法で用いるキーワード／分類テーブル作成装置および方法を提供している。

【0012】この発明によるキーワード／分類テーブル作成装置は、分類があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力するための手段、入力されたキーワードごとに、それらのキーワードに関連する分類の関連度合を求め、関連度合の大きさの順序にしたがって所定数の分類を選択する手段、キーワードごとに、それに関連する選択された分類の関連度合に基づいてキーワードを評価し、関連度合の低い分類のみが関連するキーワードを削除するキーワード評価手段、および削除されずに残ったキーワードのそれぞれについて、そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成する手段を備えている。

【0013】好ましい実施態様においては、上記キーワード評価手段は、2つの分類間の関連性の強さを表わすあらかじめ作成された分類間距離テーブルを参照してキーワードを評価し、分類間距離の大きい2つの分類が関連するキーワードを削除するものである。

【0014】この発明によるキーワード／分類テーブル作成方法は、分類があらかじめ付与された複数の分類付与済文書のそれぞれについて、それらの文書に付与された分類とそれらの文書から抽出されたキーワードとを相互に関連させて入力し、入力されたキーワードごとに、それらのキーワードに関連する分類の関連度合を求め、関連度合の大きさの順序にしたがって所定数の分類を選択し、キーワードごとに、それに関連する選択された分類の関連度合に基づいてキーワードを評価し、関連度合の低い分類のみが関連するキーワードを削除するとともに、2つの分類間の関連性の強さを表わすあらかじめ作

成された分類間距離テーブルを参照してキーワードを評価し、分類間距離の大きい2つの分類が関連するキーワードを削除し、削除されずに残ったキーワードのそれぞれについて、そのキーワードに関連の深い所定数の分類およびその分類の関連度合を対応させて記憶するキーワード／分類テーブルを作成するものである。

【0015】分類付与済文書への分類の付与は、一般に専門家によって行なわれるであろう。上記分類間距離は後述する実施例では分類間の技術距離として具体化されている。

【0016】この発明によると、あらかじめ分類が付与された分類付与済文書における分類とキーワードのデータを用いて、自動分類付与のためのデータ・ベースとなる分類間距離テーブルおよびキーワード／分類テーブルが作成されている。既存の分類付与済文書に基づいてデータ・ベースが作成されるので、上述した従来例のように人間が生成規則辞書を作成する煩わしさがなくなる。また、データ・ベースの作成のためにより多くの情報（分類付与済文書のデータ）を与えれば与えるほどより正確な分類間距離テーブルおよびキーワード／分類テーブルが得られる。すなわち、この発明は学習機能をもっており、この学習により、より正確な分類の自動付与が達成される。さらに、この発明では分類間距離という概念を導入してこの分類間距離をキーワード／分類テーブルの作成および自動分類付与処理に利用しているので、妥当でない分類の付与を排除してより正しい分類の付与が可能となる。

【0017】

【実施例の説明】図1は自動分類付与装置の電気的構成の概要を示している。

【0018】自動分類付与装置は最も好ましい形態においてはコンピュータ・システム10を含み、このコンピュータ・システム10には入力装置11、出力装置12、内部メモリ13および外部メモリ14が接続される。入力装置11は後述する分類付与済文書の分類（コード）、キーワード等を入力するとともに、分類未付与文書に記載された文章を入力するものであり、キーボード、マウス、イメージ・リーダ等を含む。分類未付与文書はキーボードから入力してもよいし、イメージ・リーダによって読込んだドット・データを文字コードに変換する文字認識処理により入力を達成することもできる。出力装置12は主に分類結果を出力するものであり、CRT表示装置やプリンタを含む。分類結果は好ましくは文書の分類欄にプリンタによって印字される。内部メモリ13はコンピュータ・システム10のプログラムを格納するとともに各種処理のためのワーク・エリア（後述する各種テーブルの作成等）を含む。外部メモリ14は入力された文書データ、分類データ等を記憶する。プログラムを外部メモリ14に格納しておいてもよい。

【0019】自動分類付与装置はあらかじめ分類が付与

された複数の文書（分類付与済文書）における分類とキーワードに関するデータに基づいて分類のための基礎データを作成し、この基礎データを用いて分類未付与文書にその記載内容に適した分類を付与するものである。分類付与処理のための基礎データとしては、分類間の技術距離テーブル（図4）とキーワード／分類テーブル（図8）とがある。したがって、自動分類付与装置は、分類付与処理（図11および図14～図16）に先だって、分類間の技術距離テーブル作成処理（図3）およびキーワード／分類テーブルの作成処理（図7）を実行する。

【0020】ここで文書とは文字で記載された内容が情報としての意味をもつすべての文書を含む。もちろん文書は人間が読むことができる形態で表わされていても、マシン・リーダブルな形態で表わされていてもよい。最も典型的な文書には技術文書があろう。中でも、特許公報、その抄録のような特許文献が最もなじみが深いものかも知れない。分類とはこのような文書を大系化して整理するためにその内容に応じて文書をグループ分けするのに用いる記号である。分類は大分類、中分類、小分類のようにヒエラルキー構造とすることもできよう。最も身近な分類には特許関係分書に付与されるIPC（国際特許分類）、各企業で付与する社内分類等があろう。文書に記載された内容の輪郭を端的に表現する用語はキーワードと呼ばれている。キーワードは一般的には文書の中で用いられる用語の中から抽出される。特許文献や学術論文ではキーワードが特定の欄に羅列して表わされている。

【0021】図2は分類付与済文書の一例を示している。

【0022】分類付与済文書には、文書を識別するための文書番号が付与されている。また文書に付与された分類を記載する分類欄と文書から抽出されたキーワードを記載するキーワード欄が設けられている。この実施例では一つの文書に最大3種類の分類が付与されるものとする。この明細書では分類（コード）をA、B、C、D、E、F、GおよびHと表現する。またキーワードをa、b、c、d、e、f、g、…等の小文字のアルファベットで表わす。一般には専門家によって分類が付与された文書が分類付与済文書となる。

【0023】まず図3から図6を参照して分類間の技術距離テーブル作成処理について説明する。

【0024】あらかじめ用意された分類付与済文書の分類欄に記載されている分類の組（一文書について最大3種類の分類からなる）が文書ごとに入力される（ステップ21）。一つの文書について分類の組が入力されるとP（I、J）テーブルおよびQ（I、J）テーブルのデータが更新される（ステップ22）。

【0025】P（I、J）テーブルは、図5に示すように、入力された分類の組の中で分類IまたはJが含まれる分類の組の数を、分類IとJのすべての組合せP

7

(I, J) (I ≠ J, I, J = A ~ H) について記憶するものである。Q (I, J) テーブルは、図6に示すように、入力された分類の組の中で分類IおよびJがともに含まれる分類の組の数を、分類IとJのすべての組合せQ (I, J) (I ≠ J, I, J = A ~ H) について記憶するものである。

【0026】分類付与済文書のすべてについて、その分類欄に記載されている分類の組の入力と、P (I, J) *

$$L(I, J) = 100 - [Q(I, J) / P(I, J)] \times \alpha \quad \cdots \text{式1}$$

αは定数である。

【0029】技術距離L (I, J) は0~100の間の値をとる。

【0030】分類Iと分類Jのすべての組合せについて式1にしたがって技術距離L (I, J) が算出され、図4に示すような分類間の技術距離テーブルが作成される。

【0031】技術距離L (I, J) は、文書が技術文書である場合に、それらに付与される分類間の技術上の関連性の近さ、または遠さを表わしている。技術距離が大きければ2つの分類間の関連性が小さく、小さければ大きい。

【0032】技術距離を一般文書についての分類間距離という概念に敷衍することができる。分類間距離は2つの分類間の関連性の近さまたは遠さを表わす。分類間距離または分類間の技術距離は式1のみならず他の演算式によっても定義することができよう。

【0033】分類付与済文書が10枚あったとして、それらに付与された分類の組が次の10個であったと仮定する。

【0034】(A, B, C), (A, B, D), (A, E, F), (B, F, G), (B, F, H), (C, D, E), (C, G, H), (C, G, I), (D, E, F), (D, G, H)

【0035】この場合に、P (A, B) = 5, Q (A, B) = 2となる。α = 100 とすると、分類AとBとの技術距離L (A, B) は式1にしたがうと、

【数2】

L (A, B) = 100 - (2 / 5) × 100 = 60 …式2
となる。この値L (A, B) = 60は単純化した一例であるから図4に示すものとは異なっている。

【0036】続いて図7から図10を参照して、キーワード／分類テーブルの作成処理について説明する。

【0037】あらかじめ用意されたすべての分類付与済文書に記載されている分類（最大3種類の分類）およびキーワードが、文書ごとに入力される（ステップ31）。後に示す自動分類付与処理と同じように、文書も入力して、入力された文書からキーワードを抽出するようにしてもよい。

【0038】分類付与済文書についての分類とキーワードの入力ごとに図9に示すようなキーワード別分類頻度

8

*テーブルおよびQ (I, J) テーブルのデータの更新が繰返して実行される（ステップ23）。これによりP (I, J) テーブルとQ (I, J) テーブルとが完成する。

【0027】分類Iと分類Jとの技術距離L (I, J) は、たとえば次式にしたがって算出される。

【0028】

【数1】

10 テーブルにおける度数（頻度）が加算される。たとえば、一文書について分類A, BおよびDとキーワードa, b, c, eおよびhが入力されたときには、キーワードa, b, c, eおよびhのそれぞれについて分類A, BおよびDの度数が+1される。すべての分類付与済文書についての分類とキーワードの入力が終了すると、キーワード別分類頻度テーブルが完成し、このテーブルに基づいて図10に示すようなキーワード別分類ヒストグラムがキーワードごとに作成される（ステップ32）。

20 【0039】このキーワード別分類頻度テーブルまたはキーワード別分類ヒストグラムは、キーワードごとに、そのキーワードと関連性がある分類についてその関連性（関係）の深さまたは強さを表わす度数から構成されている。度数はキーワードと分類との関係の深さまたは強さを表わしており、度数が大きいほど関係が強いといえる。たとえば、図10を参照して、キーワードaに最も関係が強い分類はAであり、次に分類Bが関係が強く、第3番目は分類Dである。

30 【0040】このようなキーワード別分類頻度テーブルまたはキーワード別分類ヒストグラムに基づいてキーワードの評価処理（その1）が行なわれる（ステップ33）。キーワードは特定の分類（できるだけ少数の分類）に強く関係している方が後に示す自動分類付与処理に役立つ。逆に言えば、強く関係している特定の分類が無く多くの分類に同程度に弱く関係しているキーワードは、分類付与処理のためのキーワードとして役に立たない。そこで、1または2, 3程度の少数の特定の分類に関係しているとは言いきれない役に立ちそうもないキーワードを削除するのがこのキーワード評価処理（その1）である。

40 【0041】一つのキーワードについて度数の大きいものから所定数（この実施例では3個）の分類を抽出し、その分類についての度数の和が求められ、これが所定数βよりも小さいかどうかチェックされる。たとえば、度数の高いものからn番目の分類コードの度数をδ(n)とすると（ここでnはキーワードを表わす符号とは異なり一般的な番号を表わす）、

$$\text{【数3】 } \delta(1) + \delta(2) + \delta(3) < \beta \quad \cdots \text{式3}$$

βはたとえば50

50 を満たすキーワードが削除される。

【0042】上述したキーワードaについては、度数の高い3種類の分類A、B、Dについての度数はそれぞれ80、70、10であり、これらの和は160であるから、キーワードaは削除されない。

【0043】続いて、既に作成された分類間の技術距離テーブルを参照したキーワードの評価処理（その2）が行なわれる（ステップ34）。

【0044】キーワード評価処理（その1）において削除されなかったキーワードには度数の高い3種類の分類が対応しているが、これらの3種類の分類の中に相互の関連性がきわめて低い分類対が含まれている場合には、キーワードと3種類の分類との関連性に疑問があると考えられるので、このようなキーワードが削除される。

【0045】このキーワード評価処理（その2）においては、あるキーワードについて度数の大きい3種類の分類をI、J、Kとすると、これらの3種類の分類から選択された1対の分類間の技術距離L（I、J）、L

（I、K）、L（J、K）のうち1つでもしきい値 y よりも大きいものがあれば、そのキーワードは削除される。すなわち、

【数4】 $\{L(I, J) > y\}$ or
 $\{L(I, K) > y\}$ or
 $\{L(J, K) > y\} = \text{真}$ ・・式4

であればそのキーワードは削除される。

【0046】たとえばキーワードaについては、図4の技術距離テーブルを参照すると、

$L(A, B) = 10$

$L(A, D) = 14$

$L(B, D) = 30$

であり、 $y = 40$ とすると、式4の条件を満たさないので削除されない。

【0047】このようにして2種類のキーワード評価処理（その1）（その2）において削除されずに残ったキーワードのそれぞれについて、そのキーワードに関係する度数の最も高い分類から3番目に高い分類までの重要な3種類の分類とその度数とが対応づけられることにより、図8に示すようなキーワード／分類テーブルが作成される（ステップ35）。たとえば、キーワードaについては、分類A（度数80）と分類B（度数70）と分類D（度数10）とが正しく関係するものとして対応づけられる。

【0048】図11は自動分類付与処理の概要を示している。

【0049】分類未付与文書に記載された文章が入力される（ステップ41）。上述したように、文書の文章はキーボードから入力されてもよいし、イメージ・リーダから入力されてもよい。または、あらかじめ外部メモリ14に格納しておいてこれを読み出してもよい。いずれにしても入力された文章を構成する各文字を表わすコードの列がコンピュータ・システム10内に入力され、このコード

列からキーワードを表わすコード列が抽出される（ステップ42）。入力された文章からキーワードを抽出する処理は公知であり、たとえば文章が分かち書きされ、助詞などの不要語が除かれることにより単語（主に名詞、動詞が含まれてもよい）が抽出される。この単語がここではキーワードとなる。したがって、先に説明したキーワード／分類テーブルに登録されていない単語（キーワード）が抽出されても問題は無い。キーワードの抽出処理の進行にともなって抽出されたキーワードは、図12に示すようなキーワード・リストに登録される（ステップ43）。

【0050】このようにして、入力された文章からキーワードの抽出処理、抽出されたキーワードのリストの作成が終了すると、キーワード・リストに挙げられているキーワードのそれぞれについて、リストの順番に、キーワード／分類テーブルに登録されているかどうか調べられ、登録されていればそのキーワードに対応する分類と度数が読取られ、キーワードごとに図13に示すような度数加算表に書加えられる。また、文類ごとに度数が加算される（ステップ44）。抽出されたキーワードがキーワード／分類テーブルに登録されていなければそのキーワードについては何らの処理も行なわれない。度数加算表はキーワードごとに、そのキーワードにキーワード／分類テーブルにおいて対応する分類についてその度数を記憶するとともに、分類ごとにその度数の合計を記憶するものである。

【0051】このようにして作成された度数加算表を用いて、また必要に応じて先に作成された分類間の技術距離テーブルを参照して分類決定処理が行なわれる（ステップ45）。

【0052】この分類決定処理において次の4種類の結論が得られる。

【0053】(1) 文書への既存の分類（コード）の付与（最大3種類の分類）

(2) 新しい分類（コード）の付与

(3) 検討中であることを示すコードの付与

(4) 分類不可能であることを示すコードの付与

【0054】図14から図16は分類決定処理（ステップ45）の詳細を示すものである。

【0055】まず図14において、先にステップ44で作成された度数加算表における分類ごとの合計度数を用いてヒストグラムが作成され、このヒストグラムが正規化される（ステップ51）。

【0056】図13に示す度数加算表に基づいて作成されたヒストグラムが図17に示されている。このようなヒストグラムの正規化は次式にしたがって行なわれる。

【0057】分類Iの正規化された度数を $D(I)$ とする。

【0058】

【数5】

$$D(I) = (\text{分類 } I \text{ の度数}) / \left(\sum_{J=A}^H \text{分類 } J \text{ の度数} \right) \quad \cdots \text{式 5}$$

【0059】正規化されたヒストグラムが図18に示されている。この正規化されたヒストグラムに基づいて分類の付与が行なわれる。

【0060】まず、正規化されたヒストグラムにおいて、度数の最も高い分類の度数が所定のしきい値TH1を越えているかどうかチェックされる（ステップ52）。図18に示すヒストグラムにおいては分類Dの度数がしきい値TH1を越えている。

【0061】このステップ52における判断がYESであれば次に、しきい値TH1を越えた度数をもつ分類が1つのみであるかどうかチェックされる（ステップ53）。

【0062】しきい値TH1を越えた度数をもつ分類が1つのみであればその分類が付与されることになる（ステップ54）。図18に示すヒストグラムではしきい値TH1を越える度数をもつ分類は分類Dのみであるから、このヒストグラムを生じさせた文書には1つの分類Dのみが付与される。

【0063】分類の付与は上述したように文書の分類欄に、付与されるべき分類を表わす符号もしくは記号またはコードをプリンタによって印字することによって、または文書番号に対応して分類を表示、プリント・アウトもしくはメモリに記憶することによって行なわれよう。

【0064】度数がしきい値TH1を越えた分類が1つだけでない場合には、度数がしきい値TH1を越えた分類が2つかどうかチェックされる（ステップ55）。

【0065】度数がしきい値TH1を越えた分類が2つの場合には、これらの2つの分類間の技術距離が技術距離テーブル（図4）を参照して求められ（ステップ56）、求められた技術距離が所定値よりも小さいかどうか判定される（ステップ57）。 *

$$L(A, C, D) = L(A, C) + L(C, D) + L(D, A) \quad \cdots \text{式 6}$$

【0072】他のすべての組についても同じように技術距離の合計が算出される。

【0073】続いて、このようにして算出された技術距離の合計がある所定値と比較され、その所定値よりも小さい組があるかどうかチェックされる（ステップ61）。技術距離の合計があまりに大きいということは、その組に含まれる分類の中に関連性の薄いものが含まれている可能性があるため、そのような分類の組を排除するためである。

【0074】技術距離の合計が所定値よりも小さい組が一つでもあれば、その中で技術距離の合計が最も小さい組が選択され、その組に含まれる3種類の分類が妥当なものとして該当文書に付与される（ステップ62）。

【0075】度数がしきい値TH1を越える分類が3つの場合にはその3つの分類についての技術距離の合計が算出され、この合計が所定値よりも小さければその3つ

*【0066】2つの分類間の技術距離が所定値よりも小さければ、これらの2つの分類は技術的な観点からいって比較的近いから、これらの2つの分類は妥当とみなされ、その2つの分類が該当文書に付与されることになる（ステップ58）。

【0067】2つの分類間の技術距離が所定値よりも大きい場合には、これらの2つの分類は比較的遠く、何らかの誤りを含んでいる可能性があるため分類不可能の旨が出力される（ステップ59）。この出力は、文書番号と分類不可能の旨を示す記号またはコードの表示、プリント・アウトもしくは記憶、または該当文書の分類欄への分類不可能の旨の印字によって達成される。

【0068】正規化されたヒストグラムにおいて度数がしきい値TH1を越える分類が3つ以上ある場合には、図15を参照して、これらの3つ以上の分類の中から任意の3つの分類を選択して一つの組を構成する。そして、各組ごとにその組に含まれる分類の技術距離の合計を技術距離テーブルを参照して算出する（ステップ60）。

【0069】たとえば、度数がしきい値TH1を越える分類がA、C、D、F、Gの5種類あったと仮定する。この5種類の分類の中から任意の3種類の分類が選ばれ組が構成される。生成される組は、(A, C, D), (A, C, F), (A, C, G), (A, D, F), (A, D, G), (A, F, G), (C, D, F), (C, D, G), (C, F, G), (D, F, G)の10組である。

【0070】組(A, C, D)の技術距離の合計L(A, C, D)は次式で求められる。

【0071】

【数6】

の分類が付与されることになるのはいうまでもない。

【0076】技術距離の合計が所定値よりも小さい組がない場合には、分類付与不可能の旨が出力される（ステップ63）。

【0077】正規化されたヒストグラムにおいて、度数がしきい値TH1を越える分類が存在しない場合には（ステップ52でN0）、まだ定義されていない新しい分類に振分けられる文書である可能性がある。図19は、度数がしきい値TH1を越えるものが存在しない場合の正規化されたヒストグラムを示している。

【0078】図16はこのような新分類の決定を含む処理を示すものである。

【0079】図19に示すように第1のしきい値TH1よりも低い第2のしきい値TH2があらかじめ定められている。度数がこの第2のしきい値TH2を越える分類があるかどうかチェックされる（ステップ64）。もし第

2のしきい値TH2を越える度数をもつ分類が存在しなければ分類付与不可能ということになる(ステップ63)。

【0080】しきい値TH2を越えた度数をもつ分類が一つであれば次にヒストグラム・パターン作成に移る(ステップ65)。図20に示すように、しきい値TH1とTH2との間を等分し複数(この例では5個)のランクに分ける。度数の高い方からランク1, 2, 3, 4, 5となっている。しきい値TH2を越える度数をもつ分類のうち上位複数種類(この例では5種類)の分類が選ばれ、これらの分類がどのランクに属するかが判定され、この判定結果に基づいて図21に示すようなヒストグラム・パターンが作成される(ステップ65)。

【0081】しきい値TH2を越える度数をもつ分類が5個以上無い場合にはしきい値TH2を越える度数をもつ分類のみでパターンを作成する。度数の高いものから合計5分類になるまで選択し、しきい値TH2以下のものにランク6を付与してまたはランクを付与せずにヒストグラム・パターンを作成してもよい。または分類不可能と判定してもよい。

【0082】一方、図22に示すように新分類テーブルと未定分類テーブルとが設けられている。同一のヒストグラム・パターンをもつ文書の数が所定数に達したときにそのヒストグラム・パターンに新たな分類コードが付与され、この新たな分類コードが付与されたパターンが新分類コードとともに新分類テーブルに登録される。同一のヒストグラム・パターンをもつ文書の数が所定数に達しないヒストグラム・パターンがそのパターンをもつ文書の数(出現回数: カウント)とともに未定分類テーブルに登録されている。

【0083】ステップ65で作成されたヒストグラム・パターンと同一のパターンが新分類テーブルにあるかどうかチェックされ、もしあればそのパターンに与えられた新分類が付与されることになる(ステップ66, 67, 68)。

【0084】新分類テーブルに同一パターンのものが存在しない場合には、作成されたヒストグラム・パターンは未定分類テーブルのパターンと比較される(ステップ69)。未定分類テーブルに同一のパターンがあればそのパターンのカウントが1つインクレメントされ(ステップ70, 71)、そのパターンのカウントが所定数に達したかどうかチェックされる(ステップ72)。

【0085】未定分類テーブルのあるパターンのカウントが所定数に達すると、そのパターンは新分類テーブルに移されかつそのパターンに新分類コードが割当てられ(ステップ73)、そのパターンと同一のヒストグラム・パターンを生じさせた文書に新たに割当てられた新分類コードが付与される(ステップ74)。

【0086】未定分類テーブルに同一パターンが存在しない場合には、作成されたパターンが未定分類テーブルに追加され、カウント1が与えられる(ステップ76)。この場合、およびステップ72において該当パターンのカウントが所定数に達しない場合には、その文書に検討中である旨のコードが付与される(ステップ75)。

【0087】ヒストグラム・パターンを構成する分類の数は5個に限られず、ランクは必ずしも必要ではない。要するに、ヒストグラム・パターンが類似しているかどうかを判定できるものであればよい。

【図面の簡単な説明】

【図1】自動分類付与装置の構成を示すブロック図である。

【図2】分類付与済文書の例を示す。

【図3】分類間の技術距離テーブル作成処理を示すフロー・チャートである。

【図4】分類間の技術距離テーブルを示す。

【図5】P(I, J)テーブルを示す。

【図6】Q(I, J)テーブルを示す。

【図7】キーワード/分類テーブルの作成処理を示す。

【図8】キーワード/分類テーブルを示す。

【図9】キーワード別分類頻度テーブルを示す。

【図10】キーワード別分類ヒストグラムを示す。

【図11】自動分類付与処理の概要を示すフロー・チャートである。

【図12】キーワード・リストを示す。

【図13】度数加算表を示す。

【図14】分類決定処理を示すフロー・チャートである。

【図15】分類決定処理を示すフロー・チャートである。

【図16】分類決定処理を示すフロー・チャートである。

【図17】度数加算表から作成されるヒストグラムを示す。

【図18】正規化されたヒストグラムを示す。

【図19】正規化されたヒストグラムを示す。

【図20】ヒストグラム・パターンの作成の様子を示す。

【図21】ヒストグラム・パターンを示す。

【図22】新分類テーブルと未定分類テーブルを示す。

【符号の説明】

10 コンピュータ・システム

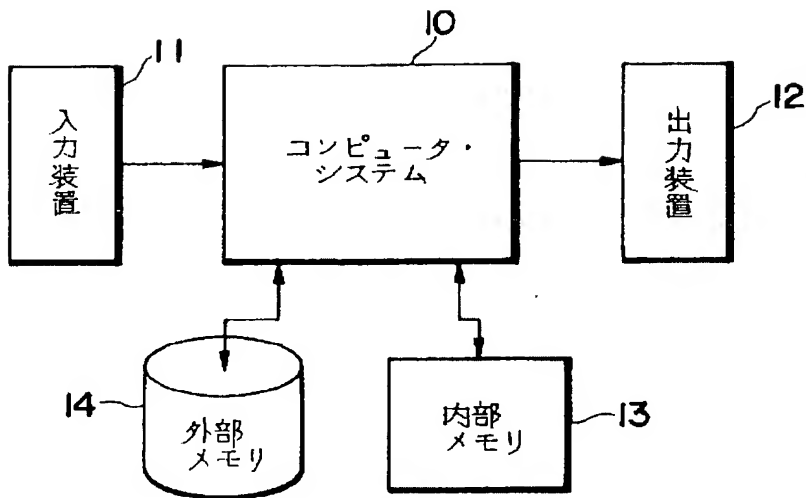
11 入力装置

12 出力装置

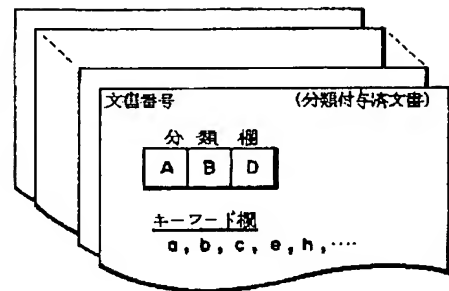
13 内部メモリ

14 外部メモリ

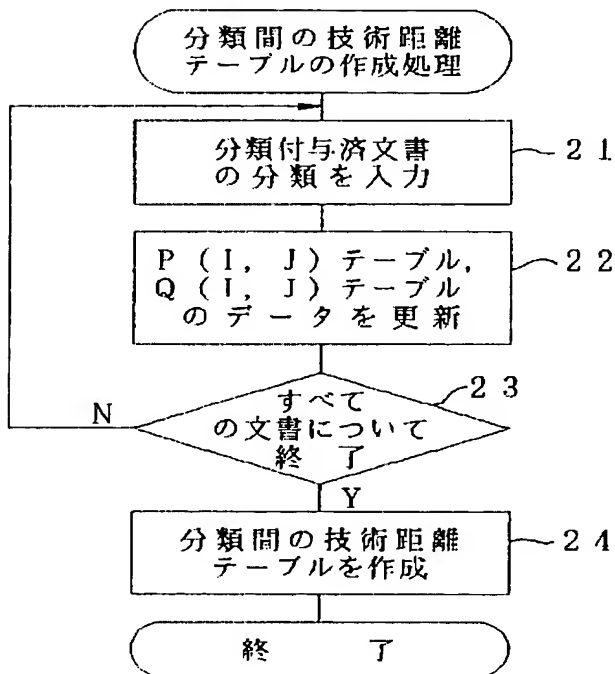
【図1】



【図2】



【図3】



【図4】

分類間の技術距離テーブル

	A	B	C	D	E	F	G	H
A		10	15	14				80
B	10		20	30				90
C	15	20		50				
D	14	30	50					
E								
F							10	20
G						10		30
H	80	90				20	30	

【図5】

P(I, J) テーブル

P(A, B) の数
P(A, C) の数
P(A, D) の数
⋮
P(B, C) の数
P(B, D) の数
⋮
P(G, H) の数

【図8】

キーワード／分類テーブル

		主 要 三 分 類 と そ の 度 数					
		分類	度数	分類	度数	分類	度数
キ ー ワ ー ド	a	A	80	B	70	D	10
	b	A	60	C	30	E	50
	⋮	⋮	⋮	⋮	⋮	⋮	
	z	E	40	G	70	H	60

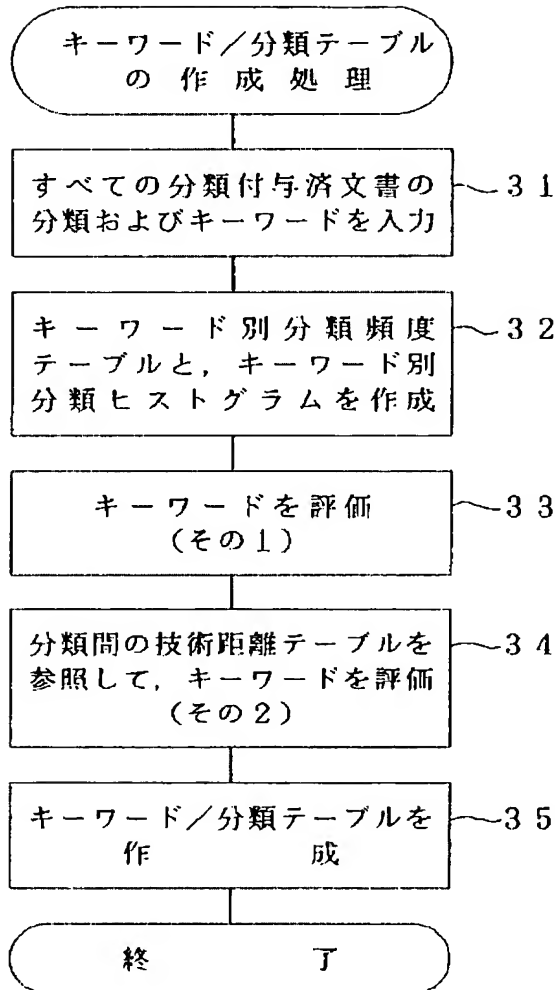
【図6】

Q (I, J) テーブル

	A	B	C	D	E	F	G	H
A								
B								
C								
D								
E								
F								
G								
H								

Q (I, J) の数

【図7】

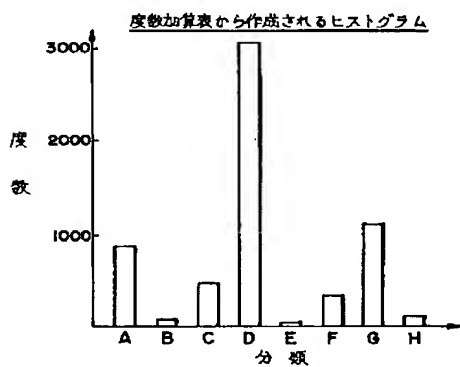


【図12】

キーワード・リスト

キーワードa
キーワードc
キーワードf

【図17】

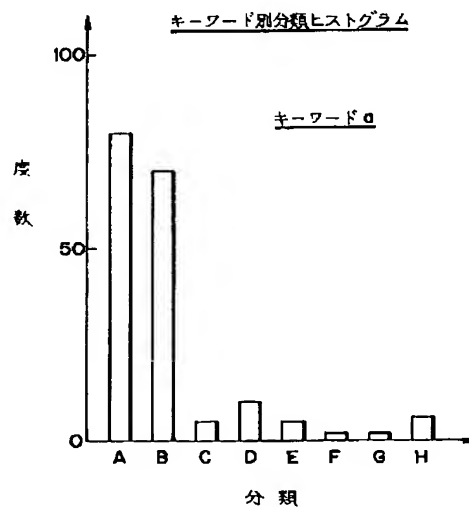


【図9】

キーワード別分類頻度テーブル

		分類							
		A	B	C	D	E	F	G	H
キーワード	a	80	70	5	10	5	2	2	6
	b								
	c								
	d								
	e								
	f								

【図10】

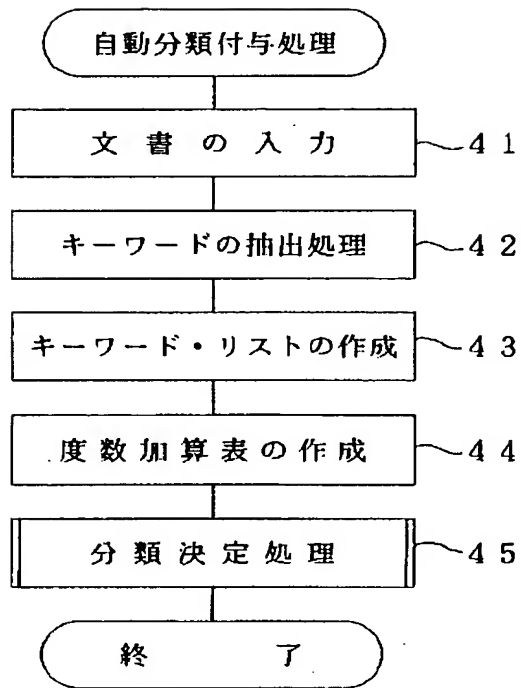


【図21】

ヒストグラム・パターン

A	2	C	4	G	4	D	5	F	5
---	---	---	---	---	---	---	---	---	---

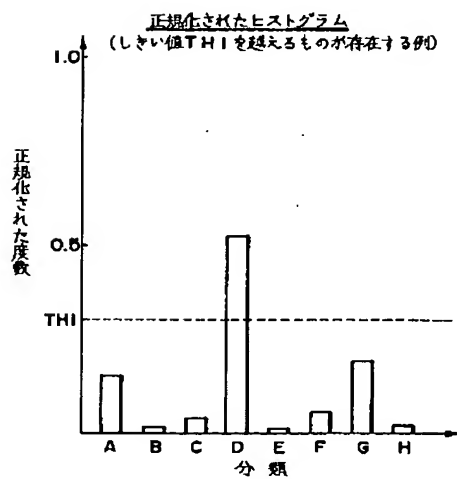
【図11】



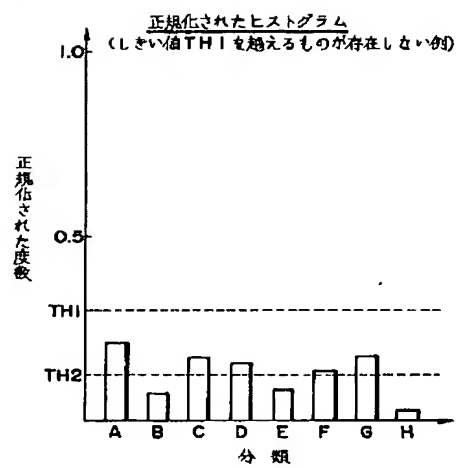
【図13】

		度数加算表							
キーワード	分類	A	B	C	D	E	F	G	H
	キーワードa	80	70		10				
	キーワードc			10	50			50	
	キーワードf	90			80		10		
	キーワードk								
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
合計		880	70	250	3050	60	330	1090	100

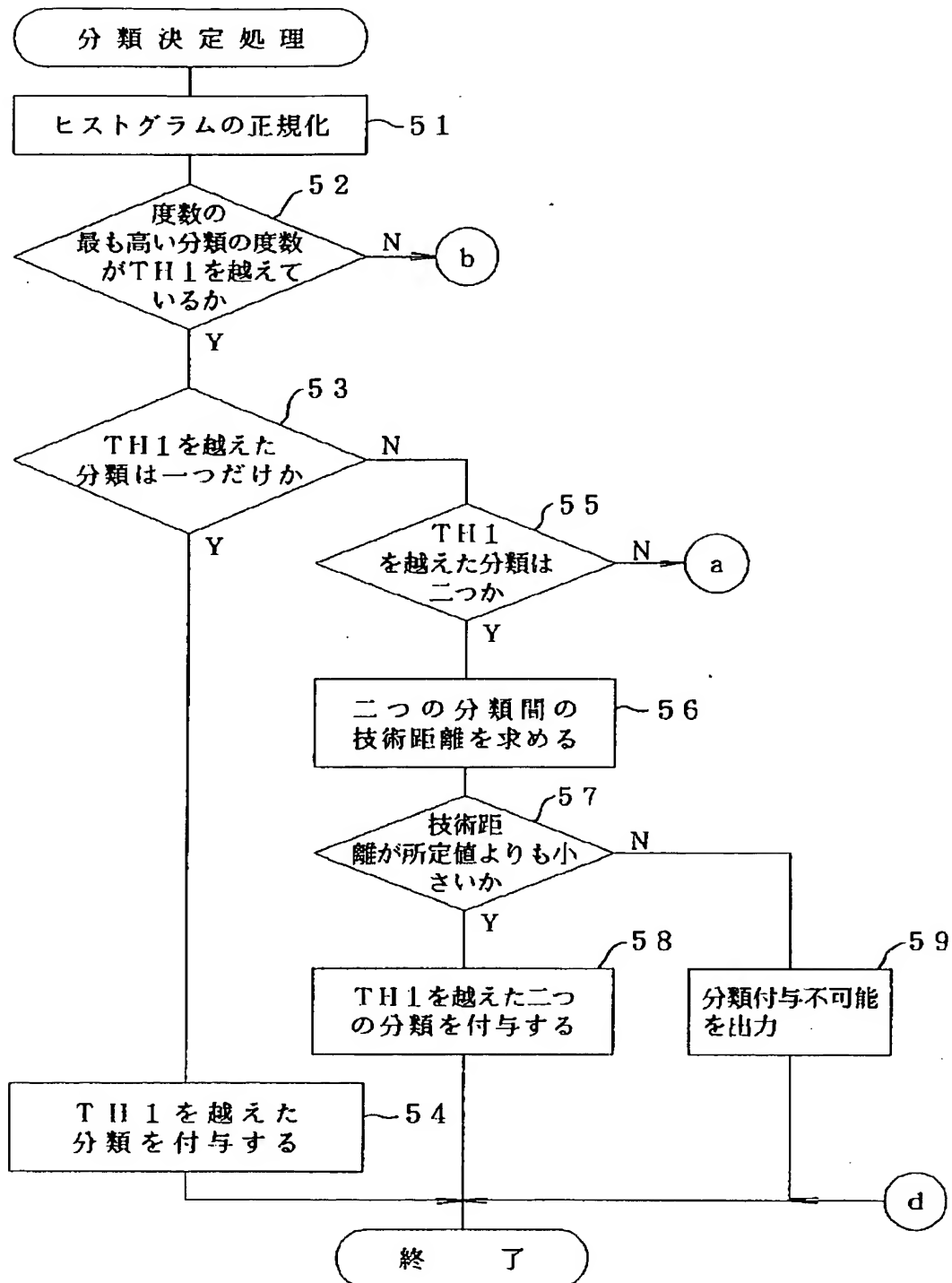
【図18】



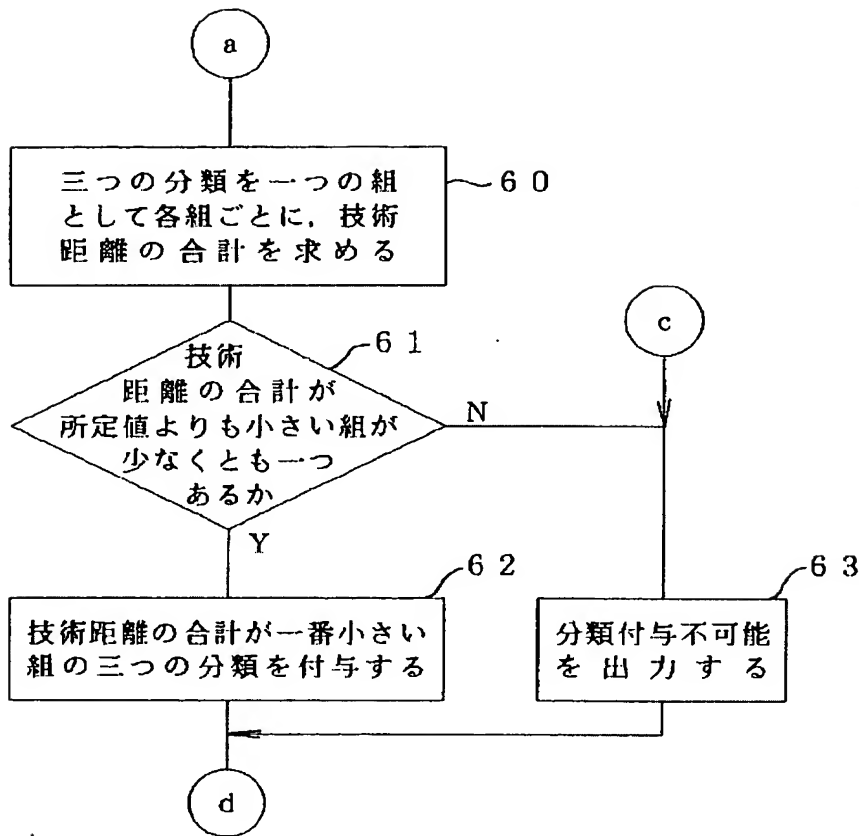
【図19】



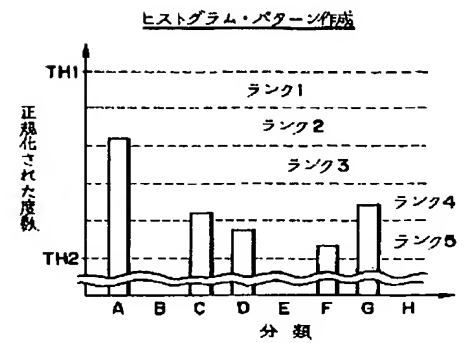
【図14】



【図15】



【図20】



【図22】

NO.	ヒストグラム・パターン									新分類コード	新分類テーブル
1	A	1	C	3	E	3				M	
2	B	2	D	2	F	1	G	1		N	
3											
NO.	ヒストグラム・パターン									カウント	未定分類テーブル
18	A	3	C	2	D	1	F	1	G	1	
19	C	2	D	2	E	2	F	1			

【図16】

